Cornell Data Science Training Program Syllabus

Spring 2017 · Gates G01 · Wednesdays from 5-6pm

This Course

When you finish this program, you will have the foundation and basic skills to contribute to any subteam in Cornell Data Science. This program introduces various machine learning models, visualization techniques, and data manipulation strategies, with applications in the R programming language. The program is open to all Cornell students at any skill level. We look forward to getting to know you over the next eleven weeks!

Course Staff

Lecturer: Dae Won Kim (dk444@cornell.edu)
TAs:

- Amit Mizrahi (am2269@cornell.edu)
- Chase Thomas (cft32@cornell.edu)
- Jared Lim (jl3248@cornell.edu)
- Kenta Takatsu (kt426@cornell.edu)

Questions In Class

- **General Questions** - If you have a question about the material which is applicable to everyone (nothing like "in my specific project") please raise your hand and we will clarify the material.
- **Personal Questions** - If you fall behind during lecture for whatever reason, or have questions specific to your own computer/project, please hold your question until after class or ask it on Piazza. Unfortunately, this may be quite a large class and we can't give people personalized help during class.
- **Course Content** - If we make a mistake on the board or say the wrong thing - don't raise your hand. Just blurt it out. Don't let us teach something that's incorrect.

Grading Policy

You have the option to enroll in the training program and get 1 credit S/U through the Cornell Data Science project team. Unless you are over credit hours, or have some conflict with your school, we will try to enroll you on student center ourselves (since add period is over).
If you are not enrolled for credit, you may still attend lectures and follow along with content on Piazza, but you will not receive feedback on your assignments. We will have 4 assignments and 2 projects throughout the course. Although it is highly recommended that you complete all the assignments, you are allowed to miss one at your discretion. Note that you cannot miss either project. We're grading mostly for completeness. If you thoughtfully answer all the questions on an assignment, or submit a functioning model for a project, you'll get credit.

Joining the Project Team

Cornell Data Science contains several subteams aside from this course. Some past subteam projects have included things like predicting formula 1 racing winners, researching misinformation spread in social networks like Reddit, and classifying sleep stages using electromagnetic signals from the brain. Please note that although we do hope you apply to these teams after the conclusion of the course, acceptance to these other CDS groups is not guaranteed. If you do well in the course, though, it certainly helps.

Piazza

- If it can be asked on Piazza, ask it on Piazza. We check Piazza as much, if not more, than we check our emails. ▯
- We will not respond to training program questions sent to cornelldatascience@gmail.com unless they are urgent. ▯
- We will not respond to questions sent to our personal emails.

Curriculum

- Week 1 - 2/15 - **Introduction to Data Science and the R language**
  - What is data science?
  - Python vs R
  - R: the good, the bad, and the ugly
  - Basic syntax, data structures, useful functions
- Week 2 - 2/22 - **Data Wrangling**
  - Improving performance with vectorization
  - Introduction to the data pipeline
  - Type checking, data structure coercion, and constraints
  - Data cleaning and imputation methods
- Week 3 - 3/1 - **Data Visualization**
  - The importance of visualizing data
  - Powerful visualizations with ggplot2
  - Heatmaps, contour maps, and mosaic plots
  - Interactive visualization with plotly
  - Brief introduction to shiny and animation
- Week 4 - 3/8 - **Linear Regression**
  - Important statistics concepts
  - Introduction to supervised learning
  - Parametric vs non-parametric learning
  - Goals of regression
  - Linear regression model
  - Interpreting linear regression output in R
- Week 5 - 3/15 - **Logistic Regression and Decision Trees**
  - Categorical Variables
  - Logistic Regression
  - The concept of sensitivity and specificity
  - ROC curves
  - Classification and Regression Trees

- Week 6 - 3/22 - **Classifiers in Supervised Learning**
    - Goals of classification
    - Bayesian classifier: the “gold standard”
    - Naive Bayes classifier
    - K-nearest neighbors (KNN)
    - Support Vector Machine
- Week 7 - 3/29 - **Clustering and Unsupervised Learning**
    - What is unsupervised learning?
    - Latent variables
    - Recommendation systems and filtering techniques
    - Hierarchical clustering and dendrograms
    - K-means clustering
    - Principal Component Analysis - dimensionality reduction
- Week 8 - 4/12 - **Model Selection and Optimization**
    - Bias and Variance
    - The tradeoff between bias and variance
    - Underfitting and overfitting
    - Feature/Subset Selection techniques
    - Regularization Methods
- Week 9 - 4/19 - **Meta-Learning and ensembles**
    - Recap of the bias-variance tradeoff
    - Introduction to ensembles
    - Bagging
    - Boosting - xgboost, ada boost
    - Stacking
- Week 10 - 4/26 - **Analyzing Text (Amit Mizrahi)**
    - Unstructured data
    - Bag of words model
    - Document-term matrix
    - Predictive coding
    - Sentiment analysis in R
- Week 11 - 5/3 - **Big Data and Conclusion**
    - Introduction to Big Data tools: Hadoop
    - Overview of the Hadoop architecture: Distributed File systems
    - The stars of yesterday: MapReduce, Pig, Tez
    - Spark: the rising star of parallel processing world